ABSTRACT

A content delivery system having m servers, $S' = \{S_1,\ldots,S_m\}$, n active customers, $C' = \{C_1,\ldots,C_n\}$, and g geographic locations, $G' = \{G_1,\ldots,G_g\}$ is disclosed, wherein $sdel_k$ is a server delay of server S_k, $ndel_{j,k}$ is a network delay observed by customers in geographic location G_j while retrieving content from server S_k, p_j is a priority value for customer C_i, c_i is a total load of customer $C_i i$, $u_{i,j}$ is a fraction of requests coming to customer C_i from region G_j, $a_{i,j,k}$ is a mapping representing a fraction of requests coming to customer C_i from region G_j that have been redirected to server S_k, and s_k represents a load capacity of server S_k. Within such a system, a method for distributing server loads includes the steps of representing an average prioritized observed response time as

$$AORT = \frac{\sum_{i=1}^{n}\sum_{j=1}^{g}\sum_{k=1}^{m} a_{i,j,k} \times u_{i,j} \times c_i \times p_i \times (sdel_k + ndel_{j,k})}{\sum_{i=1}^{n} c_i \times p_i},$$

and then generating a mapping that assigns requests from customers to a particular server while minimizing $AORT$. A heuristic algorithm is used to generate the mapping, wherein large $a_{i,j,k}$ values are assigned to small $u_{i,j}$ x $c_i \times (sdel_k + ndel_{j,k})$ values to produce a smaller overall $AORT$ value.